

When I was young I was...

- Uncertain
- Sad
- Confused
- Never understood what I was doing



**The solution
to all problems**

Do a PhD in
statistics



But...

In science:

- small sample size but large effects,
- replication crisis,
- not directly useful

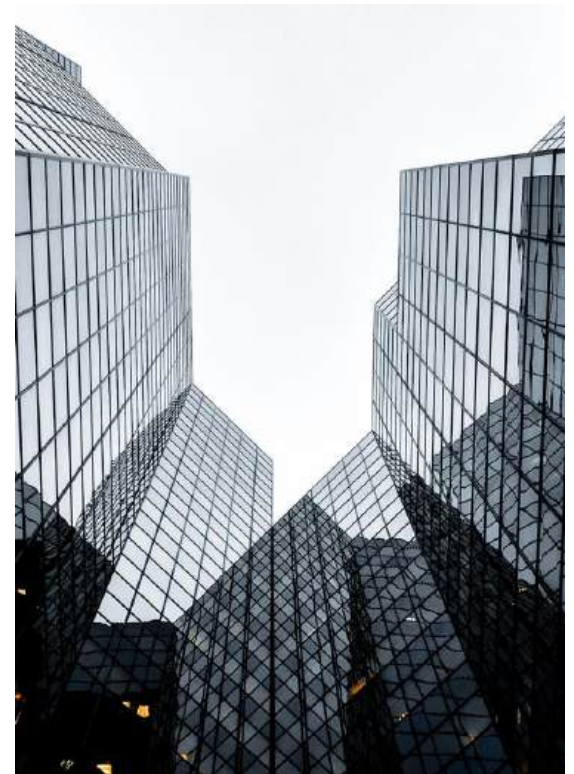
*'If I just go to business,
the sample sizes will be
huge and life will be easy'*



Where I'm coming from

Business seemed
so structured, so
organized.

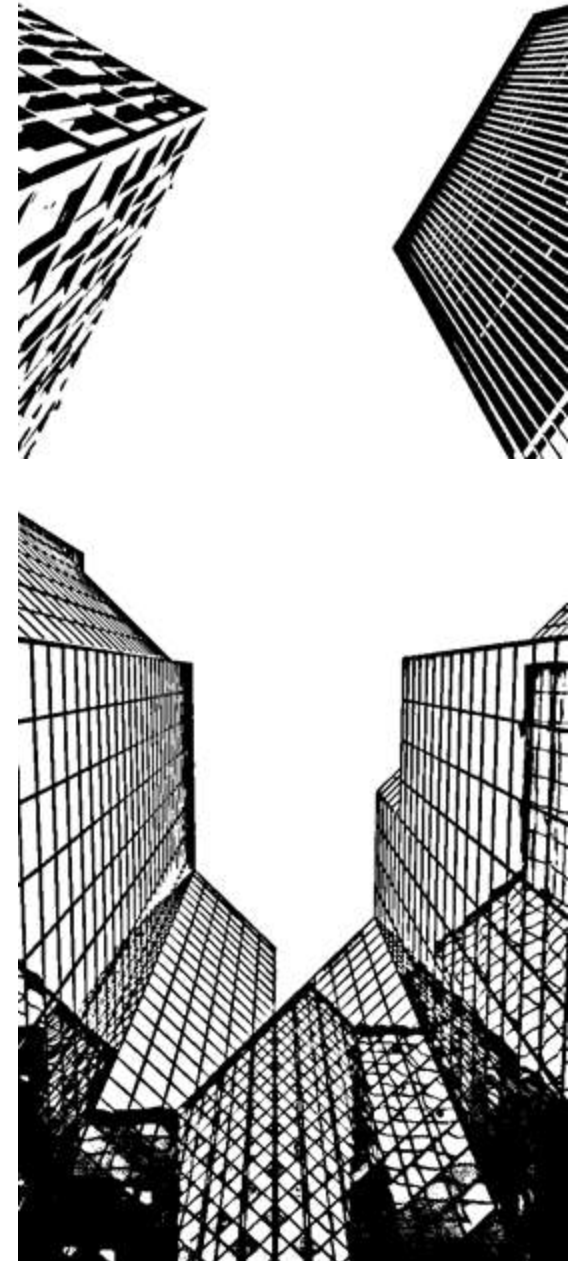
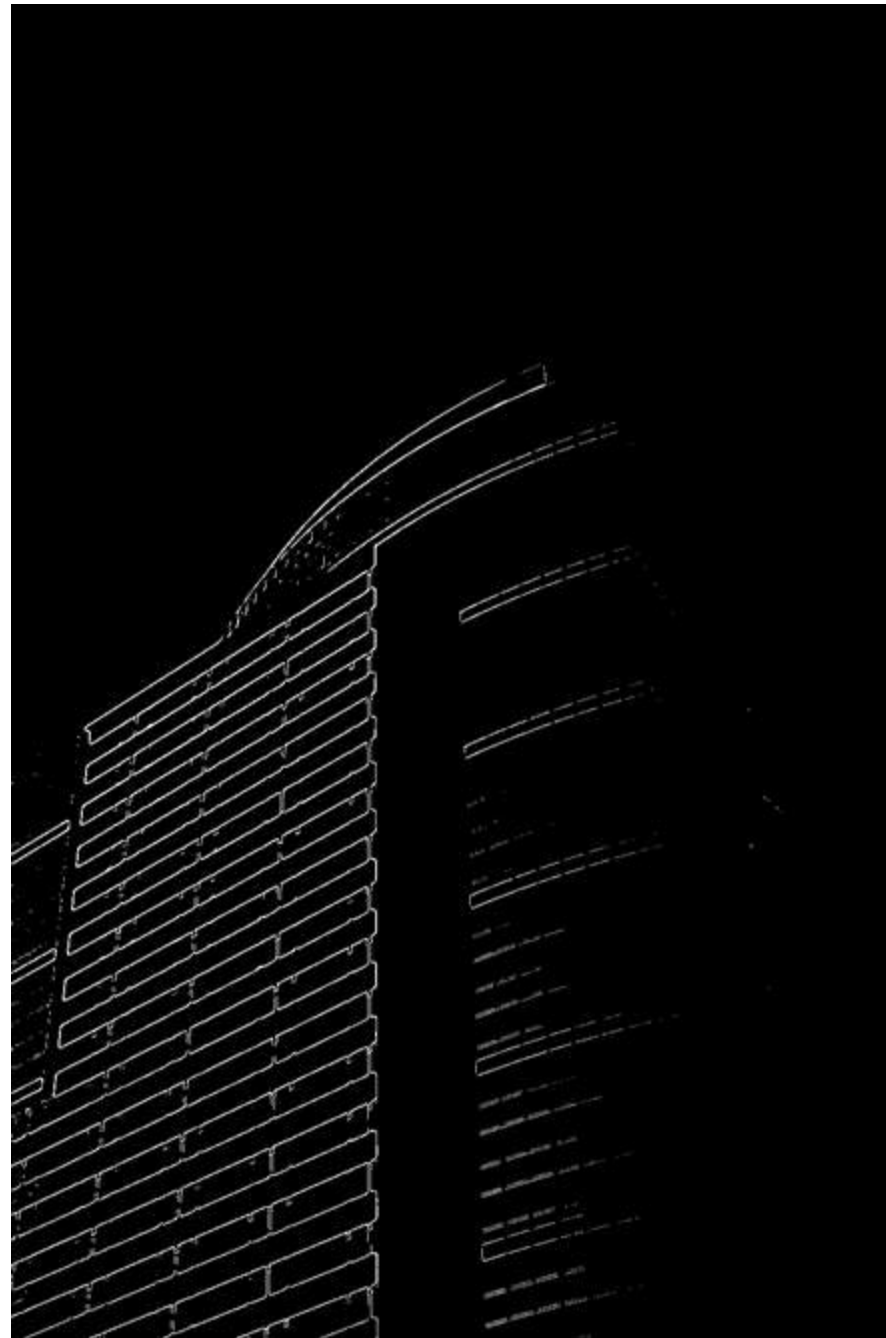
They must have
their ducks in
order...



The truth...

You can get away
with a lot of
bullshit in
business

Who's gonna call
you on it?



**Online
Experimentation
is truth**



PhD in Statistics
(Utrecht University)



Data Science
Consultant



Experimentation Data
Scientist & Statistician

First thoughts

Why simple models?!

Why is Sample Ratio
Mismatch needed?!

Variance reduction?!

Bayesian???!!!

But what works works.



Stats is about making good decisions

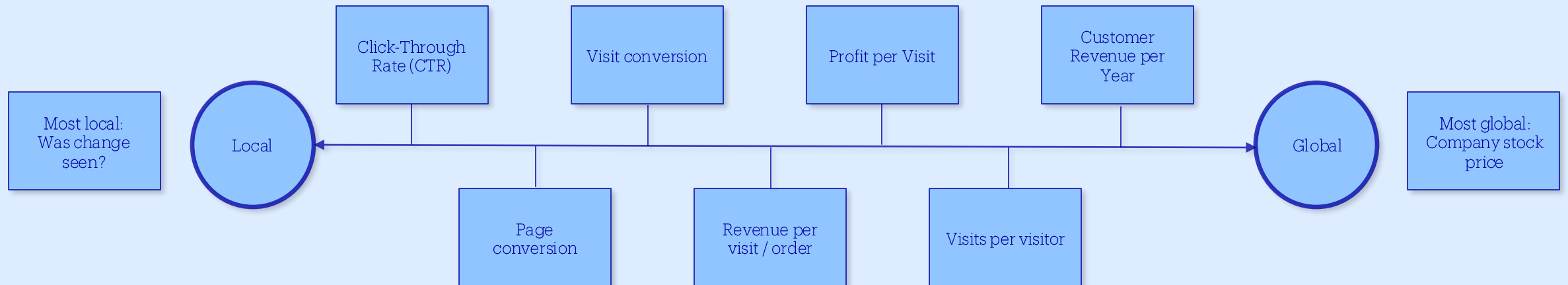
The goal should be to maximize the chance to make the right decision.

Most common mistake:

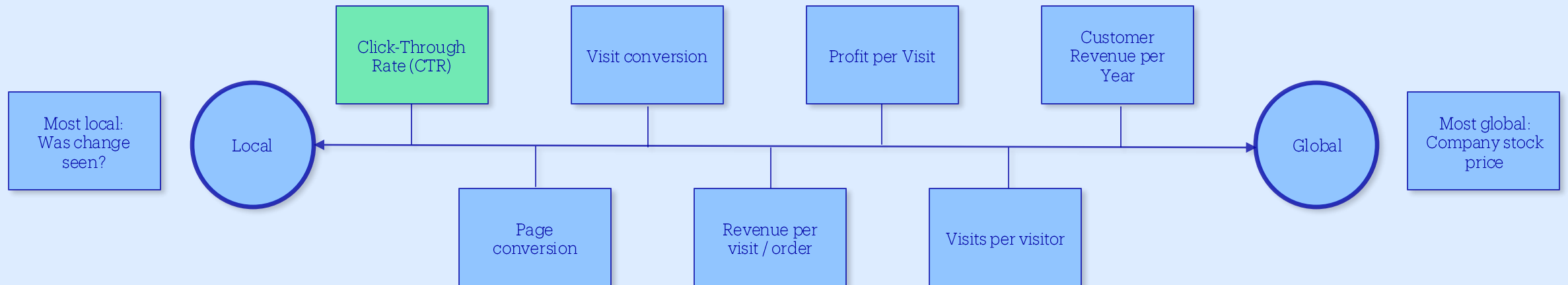
- misguided set up of metrics,
 - bad analysis,
- leading to the wrong decision.



Local vs. global metrics



Local vs. global metrics





These pixels are not available, sorry (18,18)

THIS IS WHAT CTR WANTS

Black Friday Week 20 Nov – 1 Dec

- <

Trending

Vouchers

Lightning Deals

Deals under €20

Computers & Accessories

Home & Kitchen

Electronics

Mobile Phones & Accessories

Sports & Out

>

Department

- All

Arts & Crafts

Baby Products

Beauty

Books

See more

Brands

- Logitech G

Pampers

Samsung

ECCO

See more

Customer reviews

- All

★★★★☆ & up

Price

€0 – €1,500

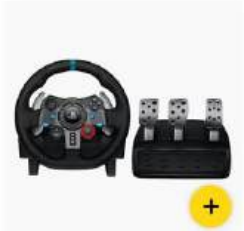


Discount

0% – 80%



Prime Programmes



57% off Limited time deal

€179⁹⁹ RRP: €419.99

Logitech G29 Driving Force Racing Wheel and Floor Pedals

Shop Logitech G deals



54% off Limited time deal

€43⁶⁹ RRP: €93.99

Pampers Baby-Dry Pants Size 8, 117 Diaper Bottoms

Shop Pampers deals



19% off Early Black Friday

€273⁹⁹ RRP: €339.99

Samsung SSD 990 PRO, 4TB, NVMe M.2, PCIe 4.0, 3% claimed

Shop Samsung deals



59% off Limited time deal

€48⁹⁵ RRP: €120.00

ECCO heren Ecco Melbourne derby veterschoen

Shop ECCO deals



53% off Limited time deal

€79⁹⁹ RRP: €169.99

Logitech G502 X PLUS LIGHTSPEED Wireless

+1 colours/patterns

Shop Logitech G deals



54% off Limited time deal

€37¹⁸ RRP: €79.99

Philips Senseo Milk Frother - Capacity of 120 ml milk

Shop Philips deals



33% off Early Black Friday

€399⁹⁸ RRP: €599.99

Nebula Mars 3 Air, Google TV Projector, Netflix Official Store

Shop NEBULA deals



35% off Limited time deal

€129⁹⁹ RRP: €199.99

Philips Airfryer with two baking drawers - Healthy, 100% oil free

Shop Philips deals



25% off Limited time deal

€74⁹⁵ RRP: €99.99

Philips 5000 Series Energy Efficient Connected Tower

Shop Philips deals



21% off Limited time deal

€22⁸⁵ Median: €28.93

Calvin Klein Boxershorts heren 3p Low Rise Trunk

+13

Shop Calvin Klein deals



32% off Limited time deal

€18⁹⁹ RRP: €27.99

Philips AquaClean Lime and Water Filter, Lime and water filter

Shop Philips deals



58% off Limited time deal

€9⁹⁹ RRP: €23.69

Page Fresh and Caring Maxi (Aloe Vera) - Wet Toilet Paper

Shop brand deals

Best-Selling Items


Within last 30 days

Within last 14 days

Within last 7 days

Filter by category

Recommended >

- 


FESTIVAL SELECTION 2025 Upgraded HY300PRO Mini Projector with WiFi and Bluetooth, Portable Native 720P, 4K Support/Decoding, Auto Keystone Correction, 180° Rotation, Smart for Home & Outdoor Use Android 11

€22.69 🔥 25K+ sold

RRP €112.16

Best-Selling Item in Office Electro... Ⓜ

★★★★☆ 1,381


Started to sell on TEMU 1 year ago
- 

FESTIVAL SELECTION 1000/1500pcs Puzzle Board with...

€23.56 🔥 6K+ sold

Best-Selling Item in Puzzles Ⓜ

★★★★★ 206

Started to sell on TEMU 3 years a...
- 


FESTIVAL SELECTION PKCELL AAA and AA Battery Set...

€12.24 🔥 97K+ sold

RRP €13.05

Best-Selling Item in Batteries & Acc... Ⓜ

★★★★★ 14,473

Brand: PKCELL
- 


FESTIVAL SELECTION 1pc Extra Large 180x200cm Hea...

€15.04 🔥 38K+ sold

RRP €24.73

Best-Selling Item in RV Furniture Ⓜ


★★★★☆ 1,360

Started to sell on TEMU 2 years ago
- 


FESTIVAL SELECTION 1pc Mattress Protector Cover, Ex...

€9.70 🔥 63K+ sold


Top Rated in Bedding Ⓜ

★★★★★ 5,402
- 


FESTIVAL SELECTION 1pc Luxury Thickened Winter Quilt - Sof...

€22.46 🔥 34K+ sold
- 


FESTIVAL SELECTION 150PSI Portable Air Compressor ...

€17.72 🔥 64K+ sold
- 

FESTIVAL SELECTION Men'S Lightweight Breathable W...

€16.20 🔥 42K+ sold
- 

FESTIVAL SELECTION 10-Pack Men'S Boxer Briefs - Br...

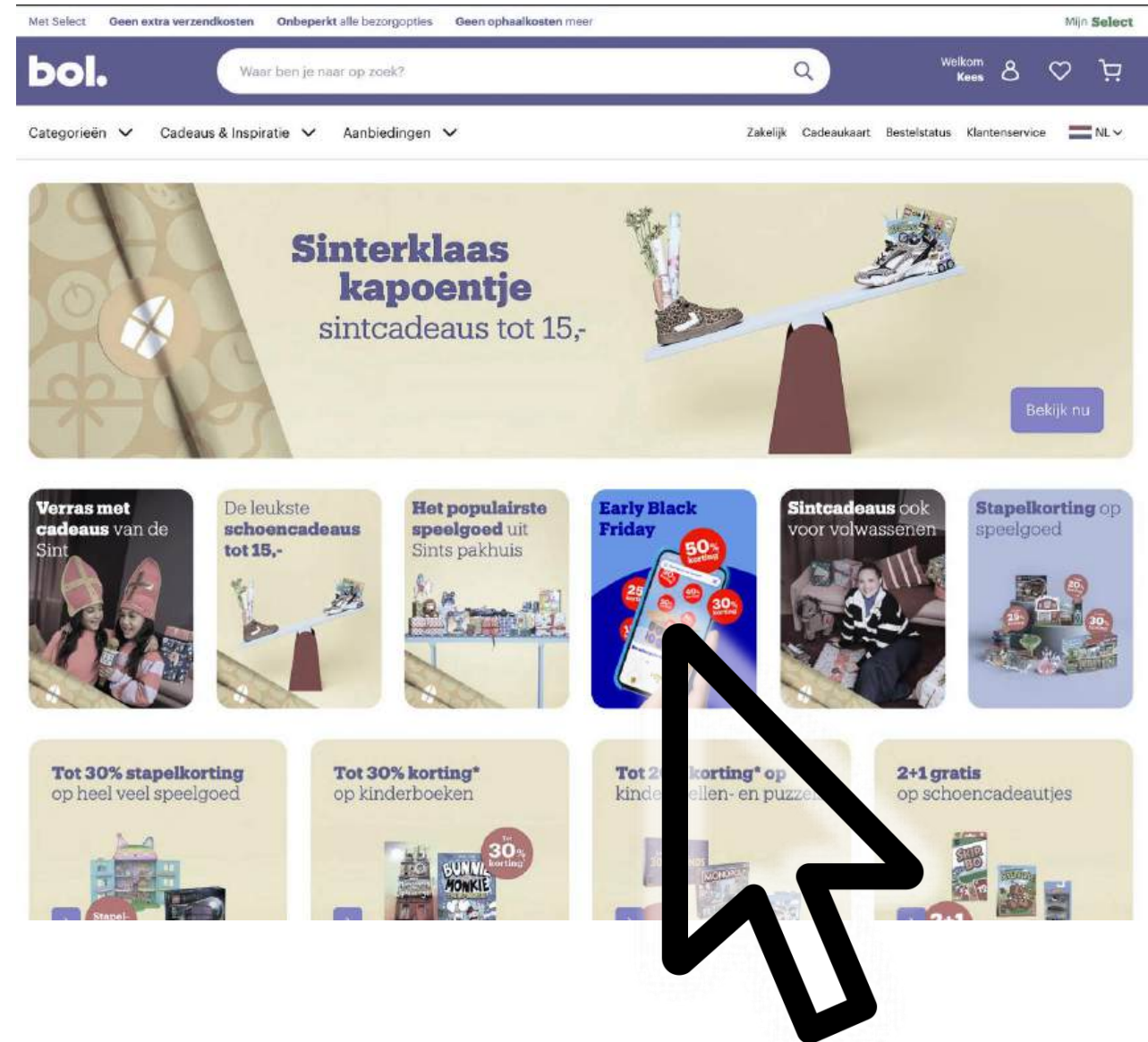
€13.43 🔥 100K+ sold
- 

FESTIVAL SELECTION Rotating Puzzle Table with 6 Lids...

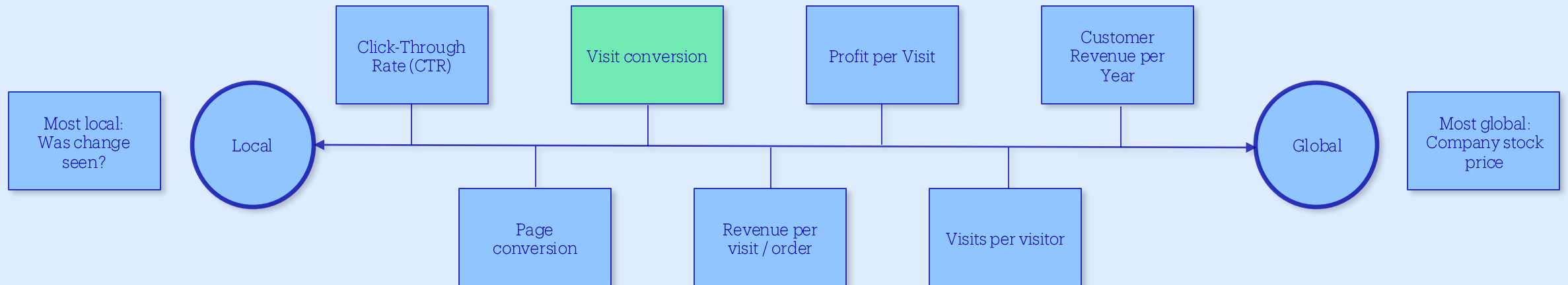
€23.31 🔥 6.8K+ sold

Click-through rate

We have a limited supply of *attention* the customer gives us.
Use it well!



Local vs. global metrics



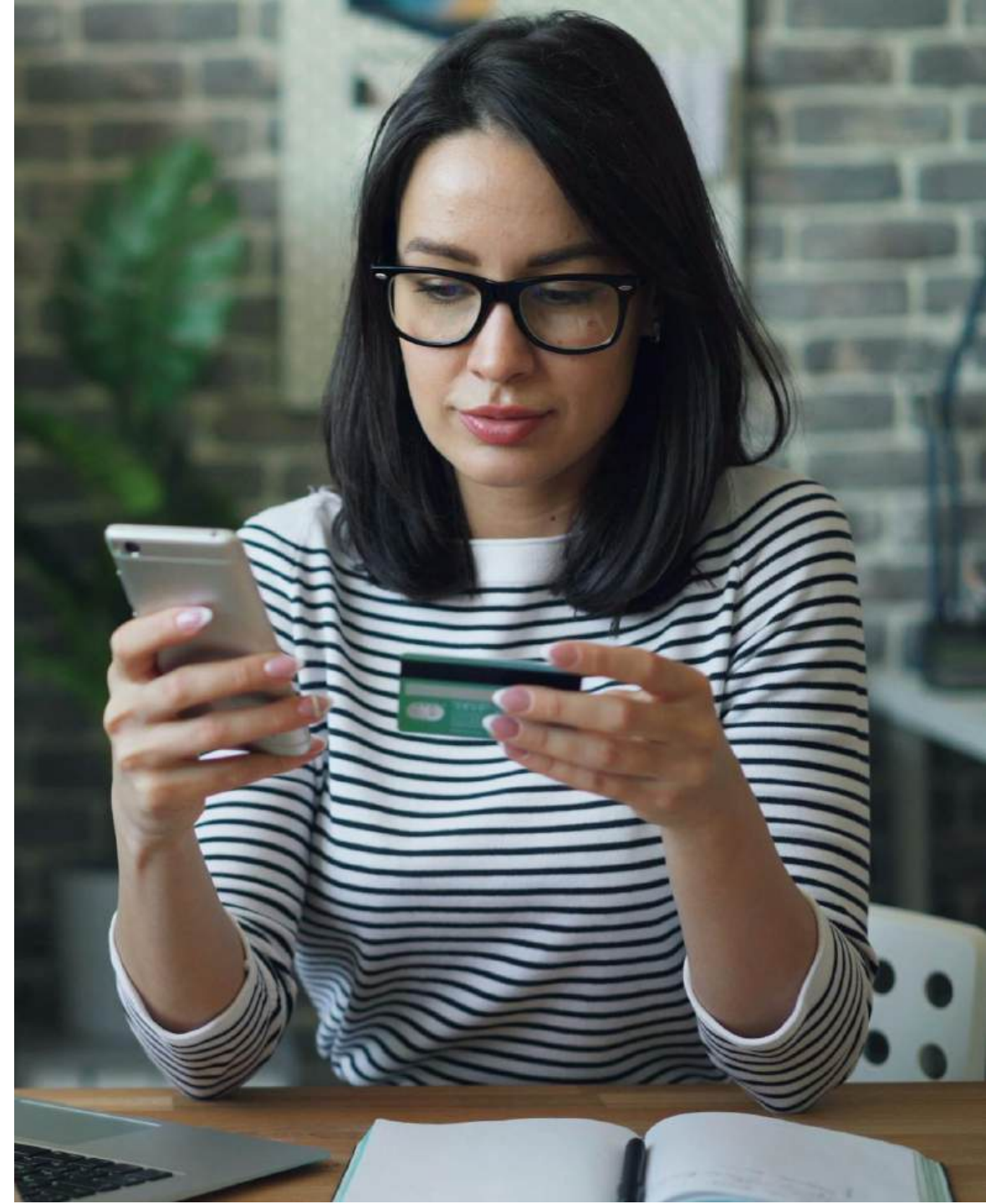
Visit conversion

Already hard to move!

- Many changes don't directly lead to sales, but are nice for the customer.

Still not always business-relevant.

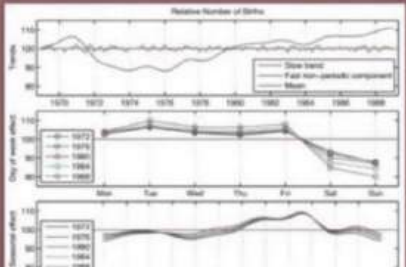
- More conversion is not always good!



Texts in Statistical Science

Bayesian Data Analysis

Third Edition




Kies je uitvoering



Hardcover
77,50

77,50

Uiterlijk 19 november in huis

Verkoop door bol

 In winkelwagen

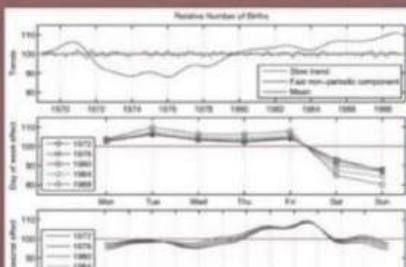
 Spaar **15 punten** 

‘To basket’

Texts in Statistical Science

Bayesian Data Analysis

Third Edition



Kies je uitvoering



Hardcover
77,50

77,50

Uiterlijk 19 november in huis

Verkoop door bol

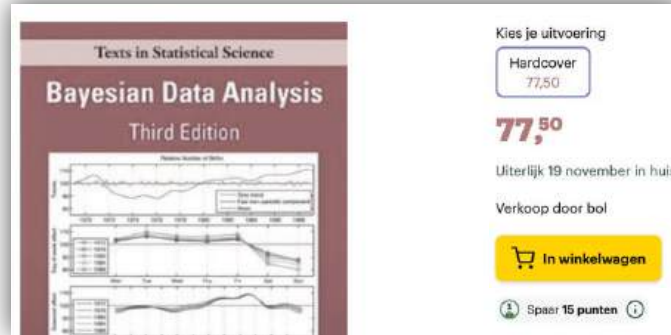
[> Naar de kassa](#)

 Spaar **15 punten** 

‘Pay now’

Conversion

Skip basket
experiment
leads to severe
order splitting.



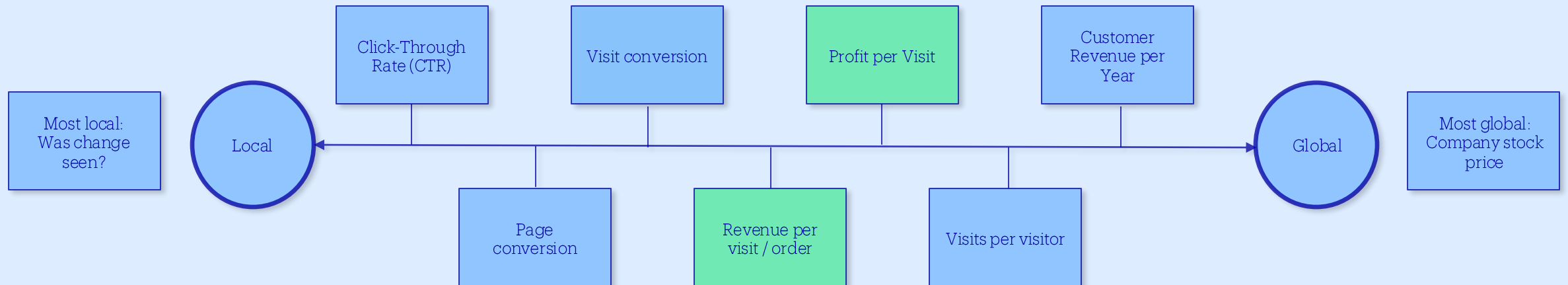
1 Conversion

Average Order
Value: €26.09

2 Conversions

Average Order
Value: €13.05

Local vs. global metrics



Revenue or profit metrics

Great for
balancing
everything.

Never*
significant.

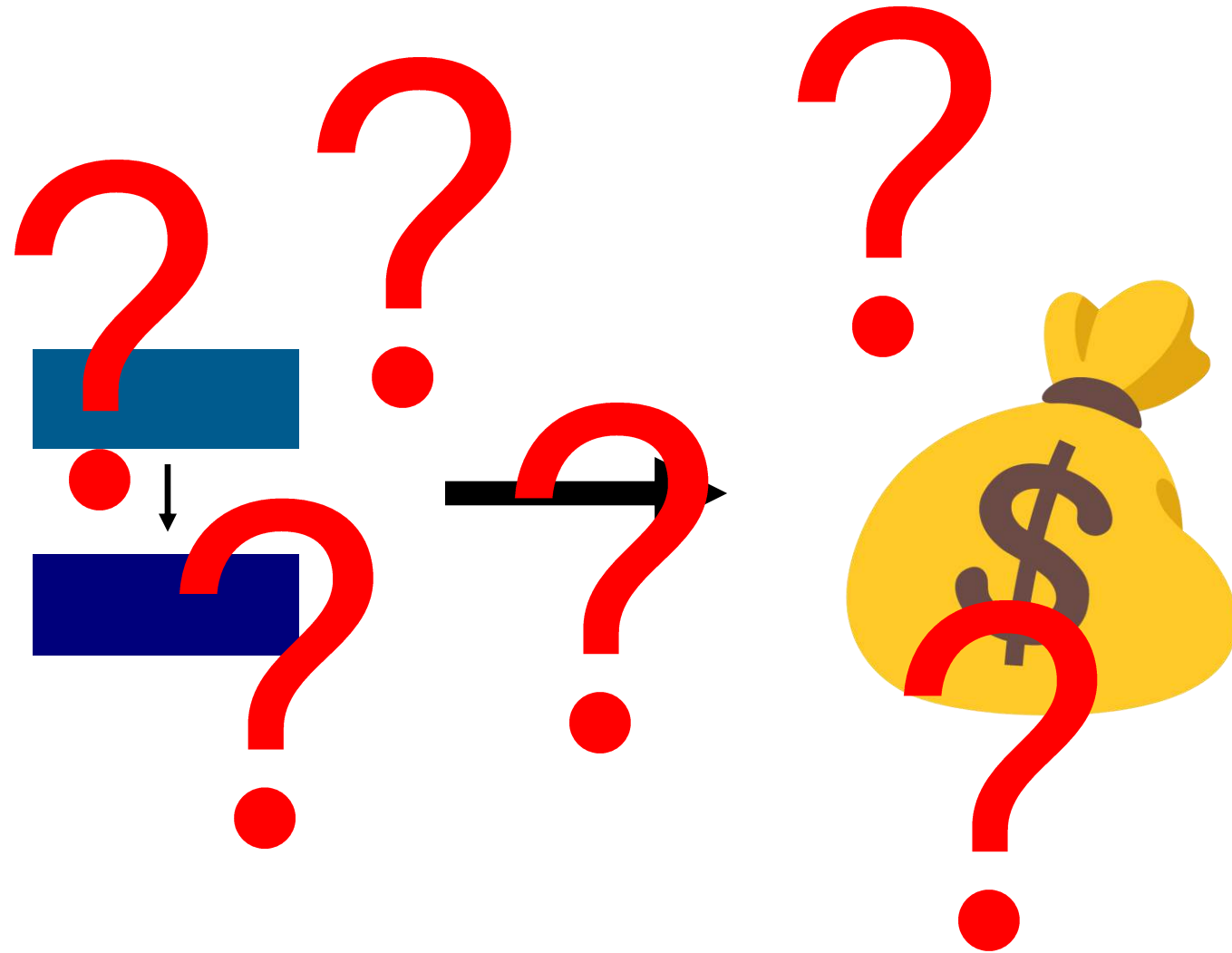


Why never significant?

Way too much noise:
The 'ten laptops' problem.

But also: suboptimal
statistical models (t-test,
Mann-Whitney U-test).





Recommendations

1. Collect your global metrics (ideally profit).
2. Balance local vs. global.
3. Check relationship to global metrics.
4. Use mutually exclusive metrics.
5. Combine or decompose metrics.



**Recommendation 1:
Get profit?**



Recommendation: Get profit

Without collecting the data, you assume 0% chance of effect in uncollected metrics.

At Bol, Profit per Visit is called Real Estate Yield, combining:

- Conversion
- Order Value
- Margin
- Ad gains
- Returns (estimated)
- Customer service calls



Recommendation 2: Balance local vs. global



Recommendation:

Balance local & global

Take the most relevant/global primary metric that you have enough data/effect size for.

Not having enough data is OK.

Use a more local metric if you must.

Never taking a decision is worse!

But think if assumptions make sense.

'Could there be order splitting?'

Then, check more global metrics.



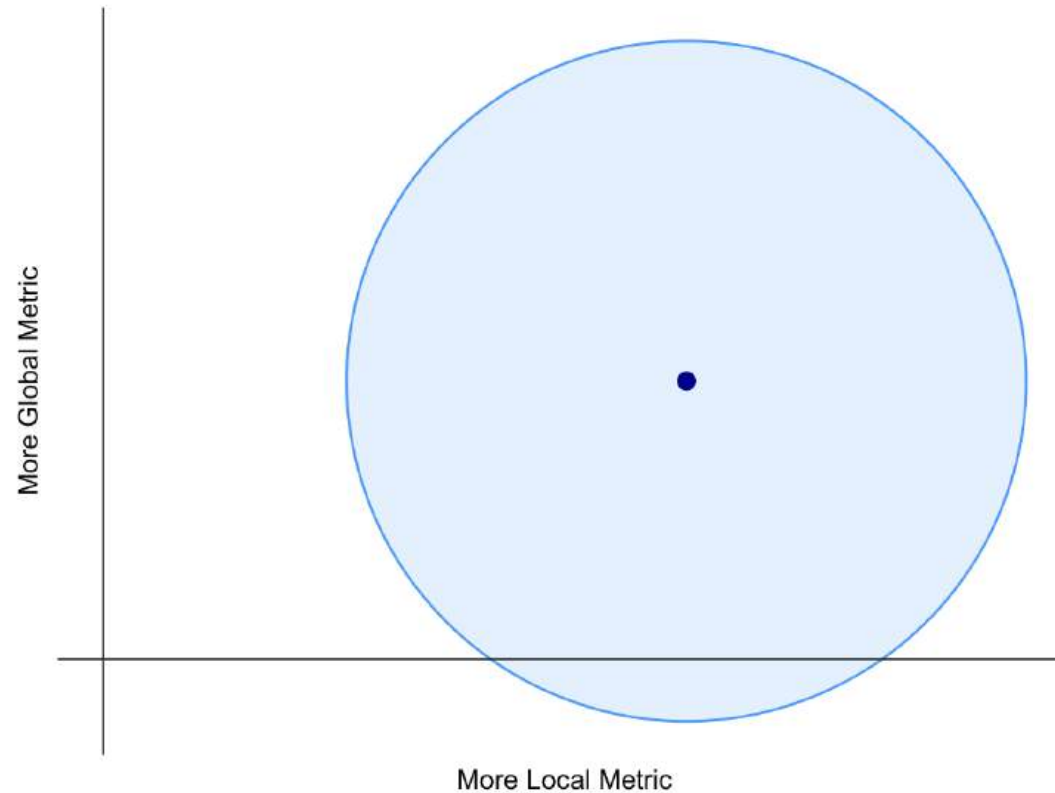
Recommendation 3: Check yourself



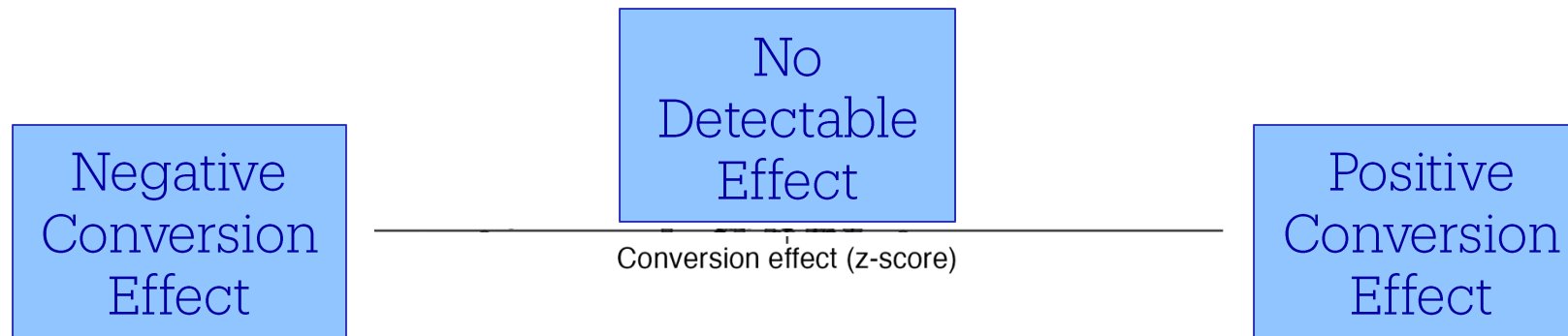
Checking your local metrics

When you get home:

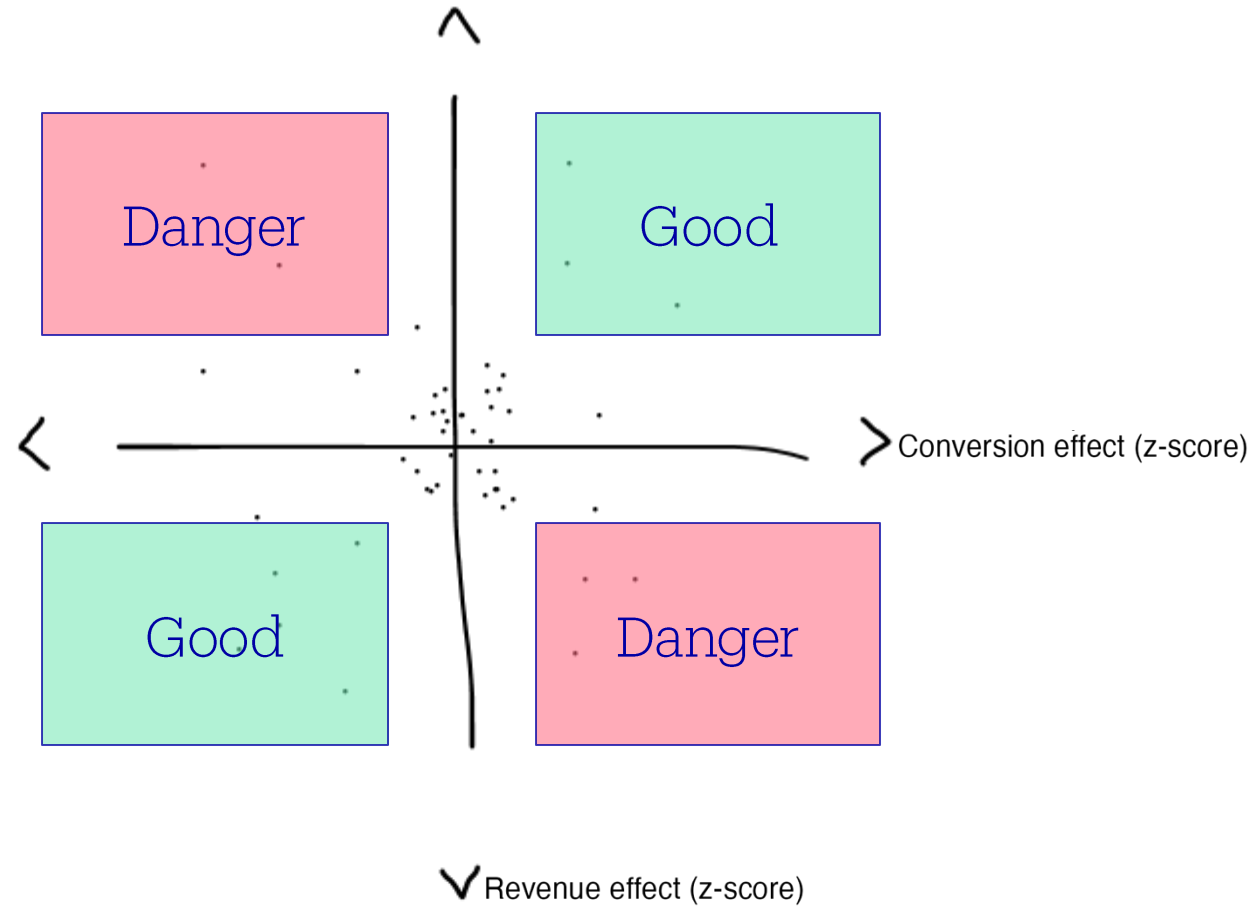
1. Collect historical data of all experiments!
2. Make a plot with your more sensitive local metric and the global metric you care more about.
3. Check: does local predict global?



Checking your local metrics



Checking your local metrics



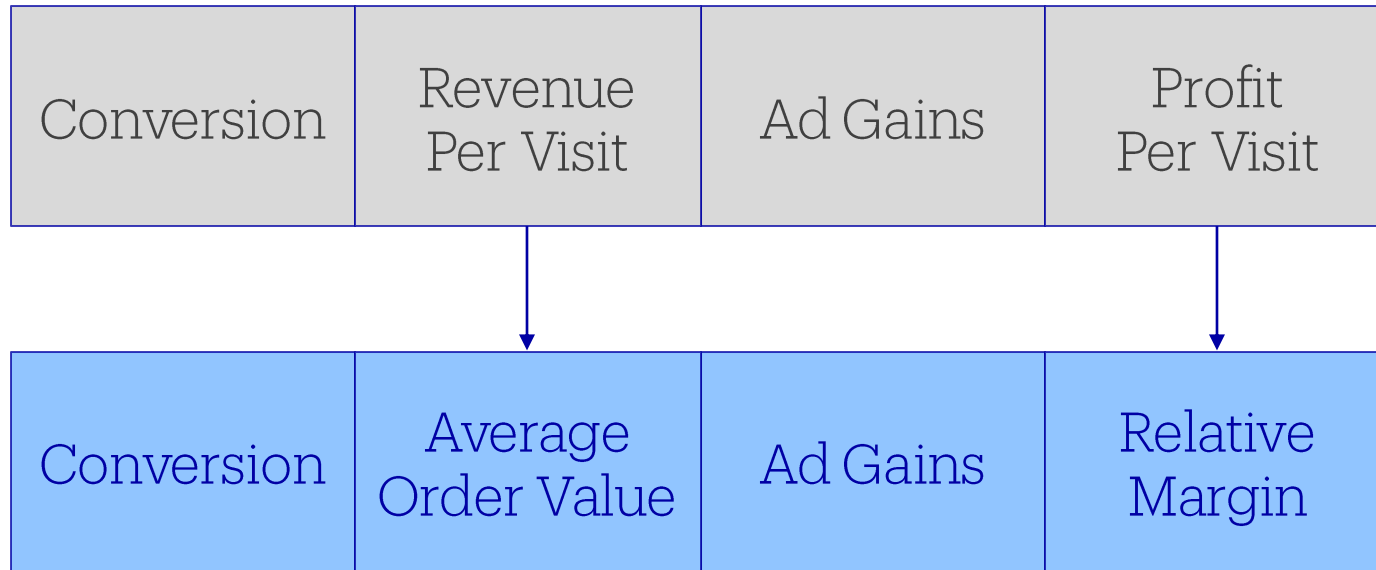
Recommendation 4: Mutually exclusive metrics



Conversion	Revenue Per Visit	Ad Gains	Profit Per Visit
------------	----------------------	----------	---------------------

These metrics are overlapping!

Conversion		
Revenue Per Visit		Ad Gains
Profit Per Visit		



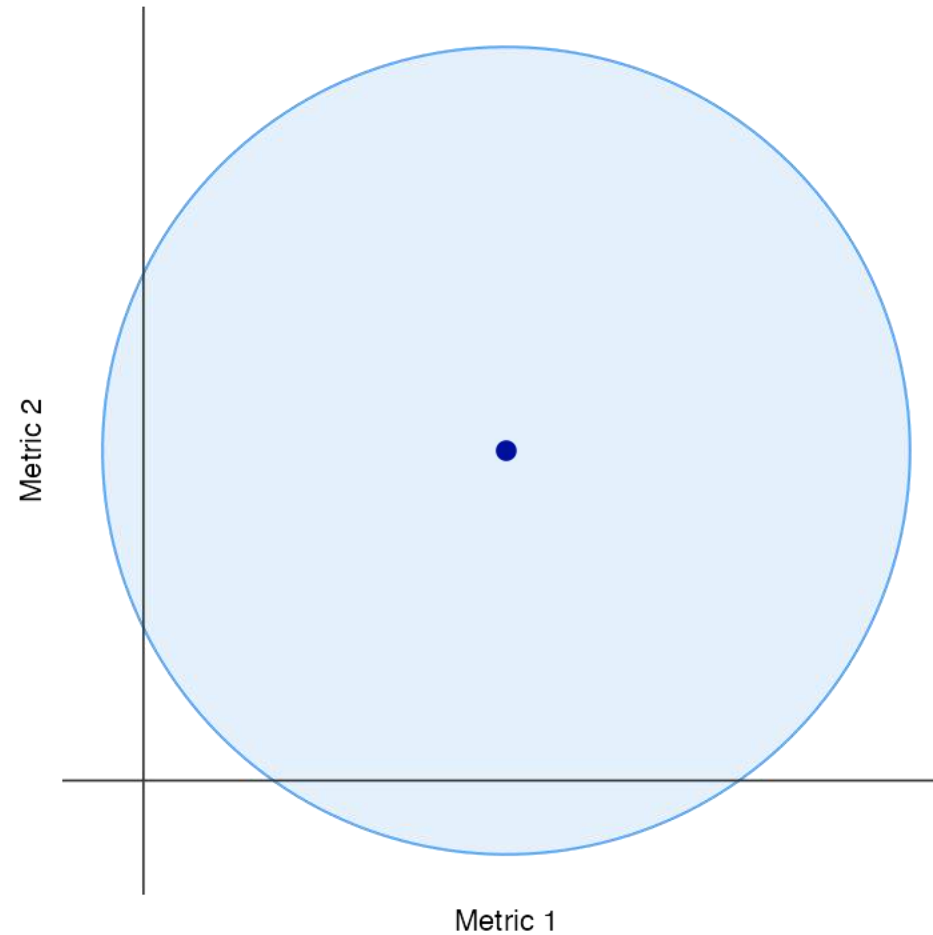
Try to look at mutually exclusive parts!

$$Profit = Conversion * Average Order Value * Margin + Ad Gains$$

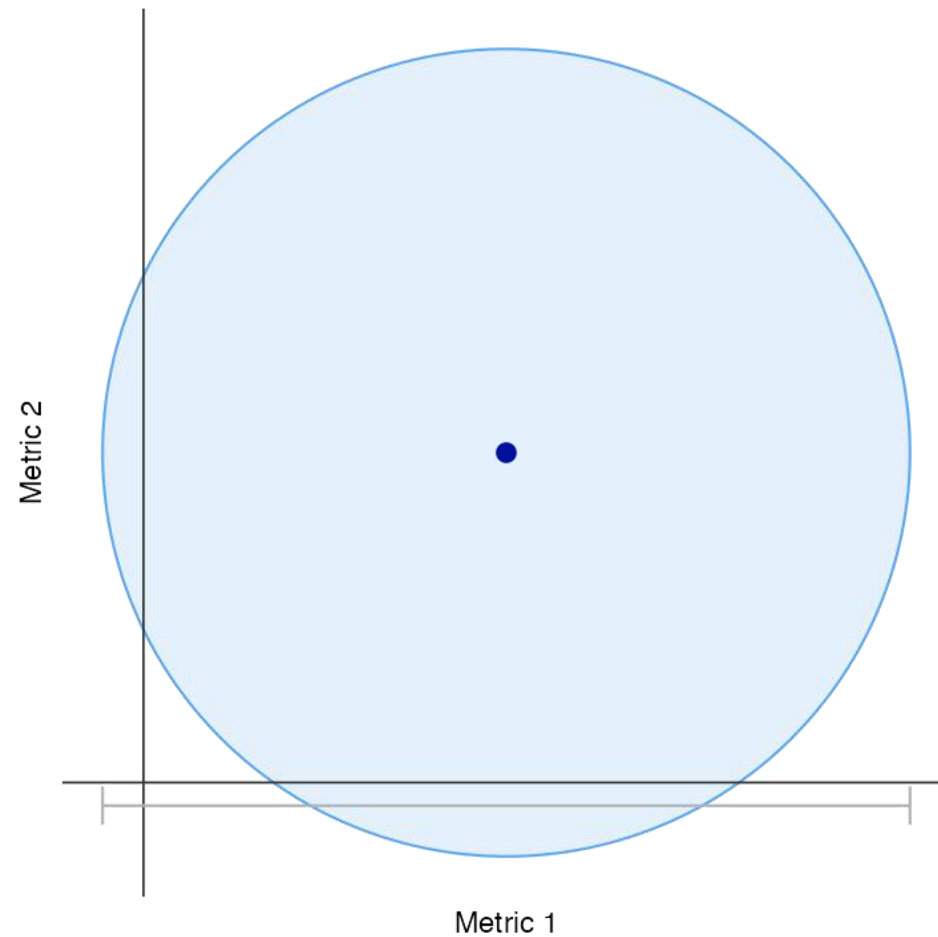
Recommendation 5: Combining and Decomposing



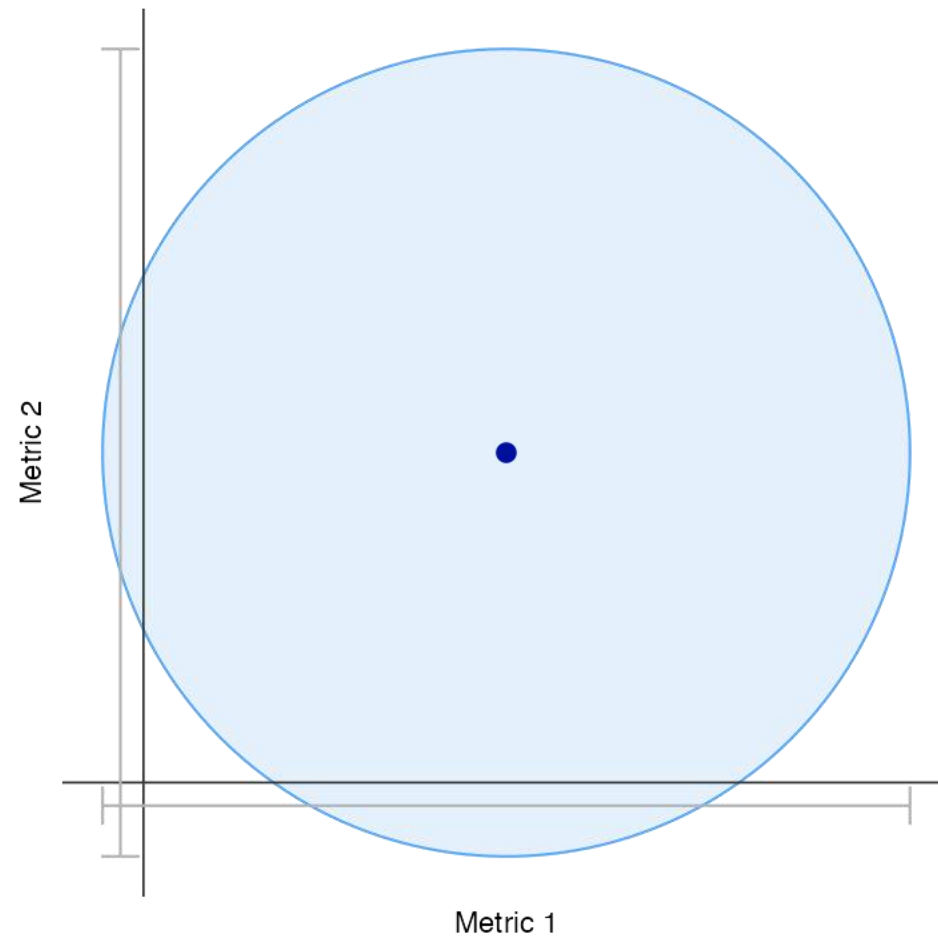
Maybe we can combine the metrics?



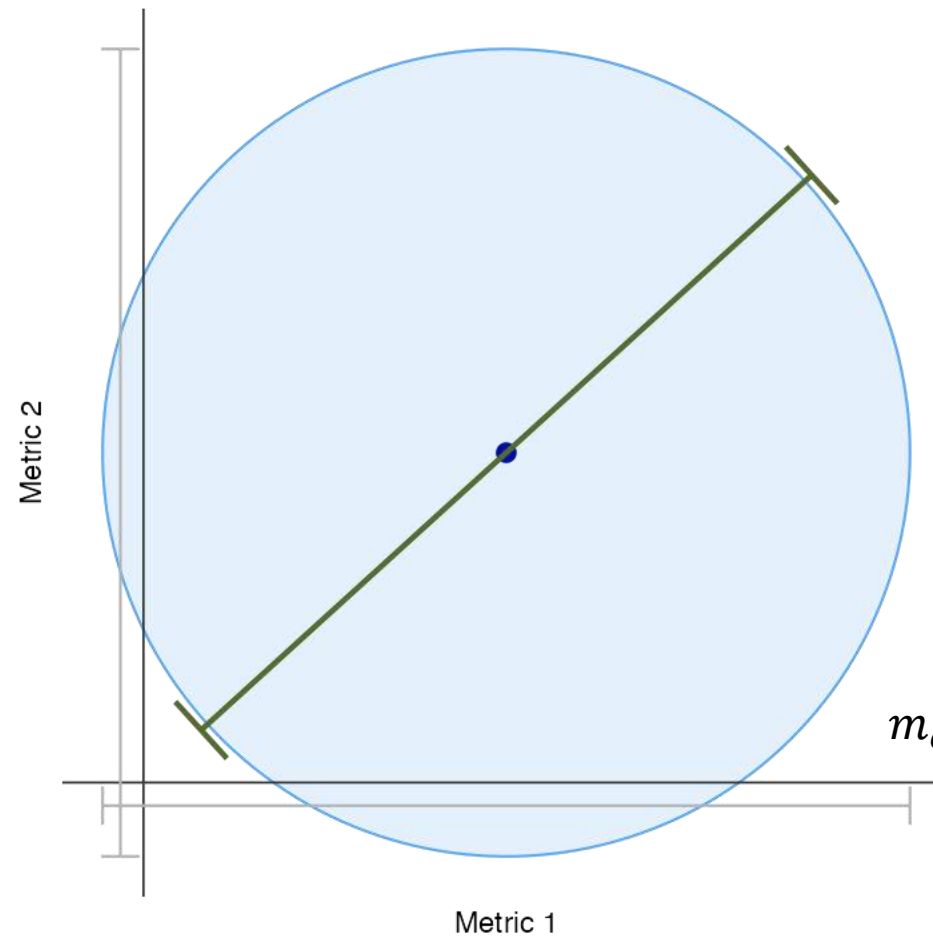
Maybe we can combine the metrics?



Maybe we can combine the metrics?



Maybe we can combine the metrics?



$$m_c = w_1 m_1 + w_2 m_2$$
$$m_c = 45\% \text{ Conversion} + 55\% \text{ AOV}$$

Metric decomposition

Metric Decomposition in A/B Tests

Alex Deng*
Airbnb
Seattle, WA, USA
alex.deng@airbnb.com

Luke Hagar
University of Waterloo
Waterloo, ON, Canada
lmhagar@uwaterloo.ca

Nathaniel T. Stevens
University of Waterloo
Waterloo, ON, Canada
nstevens@uwaterloo.ca

Tatiana Xifara
Airbnb
San Francisco, CA, USA
tatiana.xifara@airbnb.com

Amit Gandhi†
University of Pennsylvania
Philadelphia, PA, USA
agandhi@upenn.edu

Abstract

More than a decade ago, CUPED (Controlled Experiments Utilizing Pre-Experiment Data) mainstreamed the idea of variance reduction leveraging pre-experiment covariates. Since its introduction, it has been implemented, extended, and modernized by major online experimentation platforms. Despite the wide adoption, it is known by practitioners that the variance reduction rate from CUPED utilizing pre-experimental data varies case by case and has a theoretical limit. In theory, CUPED can be extended to augment a treatment effect estimator utilizing in-experiment data, but practical guidance on how to construct such an augmentation is lacking. In this article, we fill this gap by proposing a new direction for sensitivity improvement via treatment effect augmentation whereby a target metric of interest is decomposed into components with high signal-to-noise

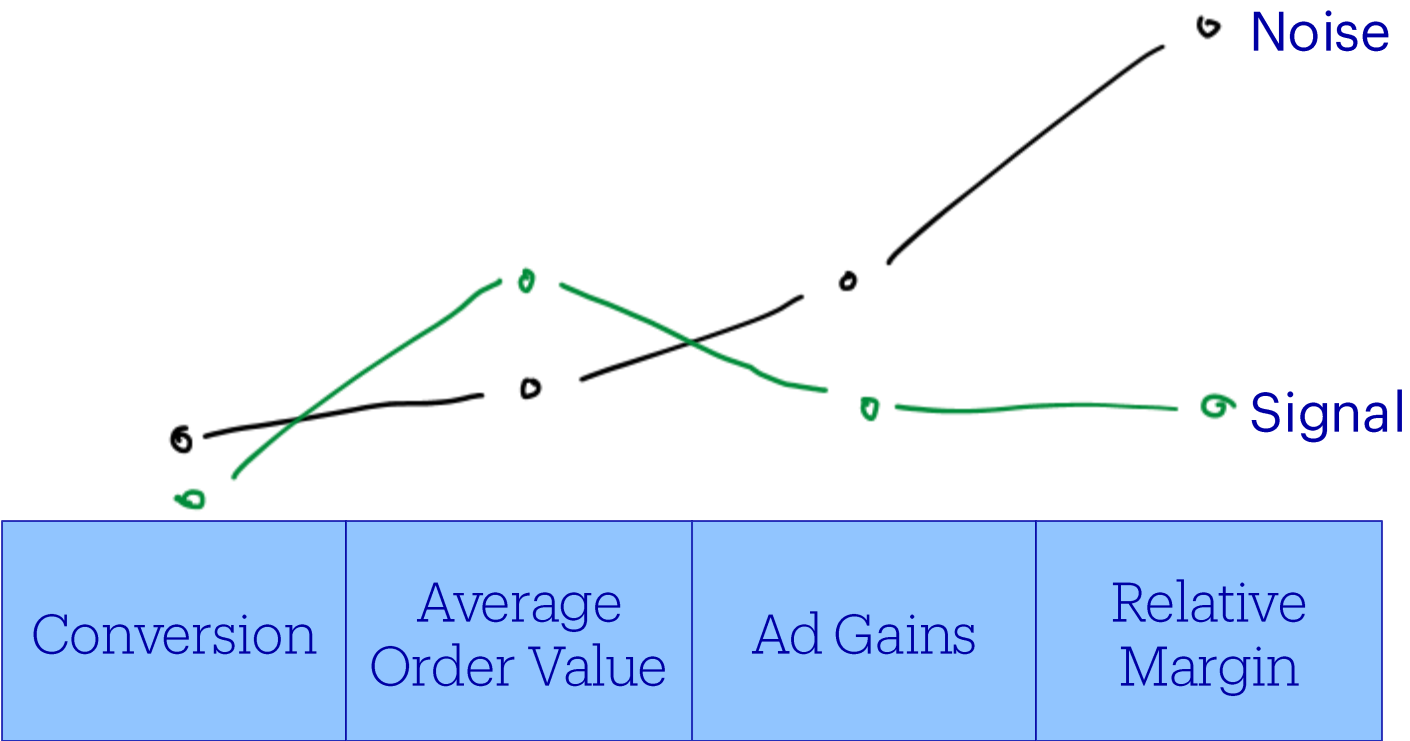
Mining (KDD '24). ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3637528.3671556>

1 Introduction

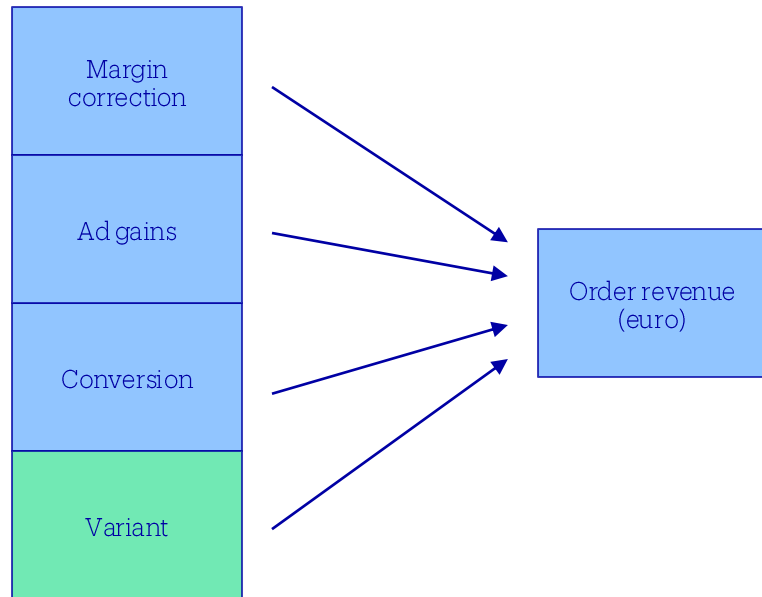
Online controlled experiments, also referred to as “A/B tests”, are an invaluable tool used by companies to test and evaluate changes to their online products. With respect to some metric(s) of interest, these experiments facilitate causal conclusions about the efficacy of such changes. Large tech companies collectively run tens of thousands of these experiments each year, engaging millions of users [23].

An A/B test typically compares two versions of a product: a new *treatment* version to the existing *control* version. Interest lies in understanding the *treatment effect* δ , which quantifies the potential improvement (with respect to some metric of interest) induced

Can we assume some of these are zero?



Predict the goal outcome with the others



Proven to be more powerful
(probably more than CUPED!).

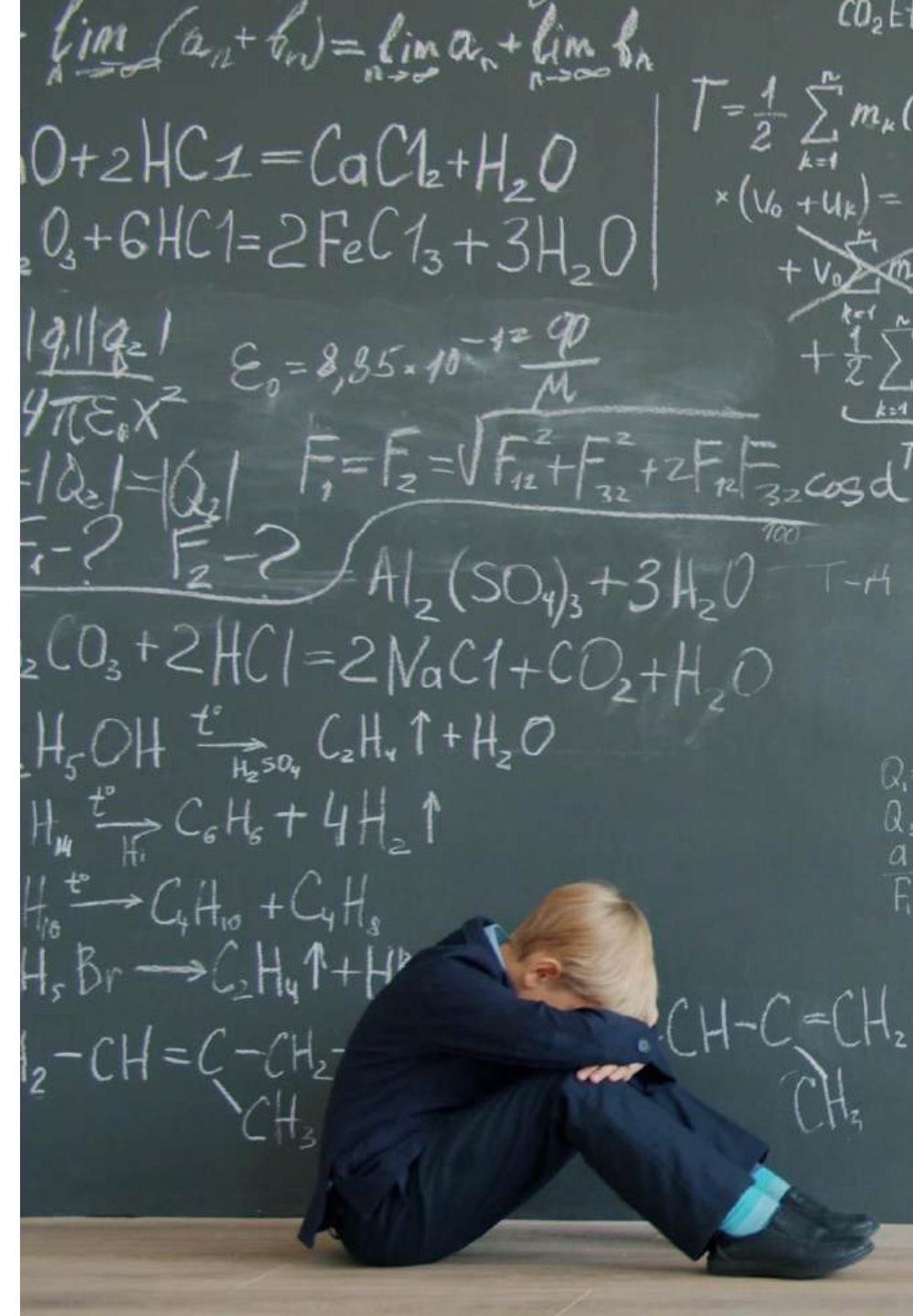
Model is a bit more complex,
but you can do it!

Note: outcome metric still
needs to be relevant!

Looking back

Okay, maybe you don't really need to do a PhD to do run a good experiment.

But strengthening your methods can really put the mind at ease.



Looking back

I'm having a great time!

Most experiments still
use bad metrics and
mediocre analysis...

Always fun stuff to do!



